Recurrent Neural Networks

Prof. Giuseppe Serra

University of Udine

Recurrent Neural Networks

Next-character prediction

Chatbots

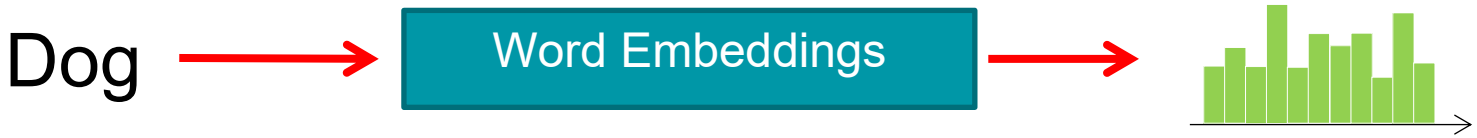
Music composition

Image captioning

Speech recognition

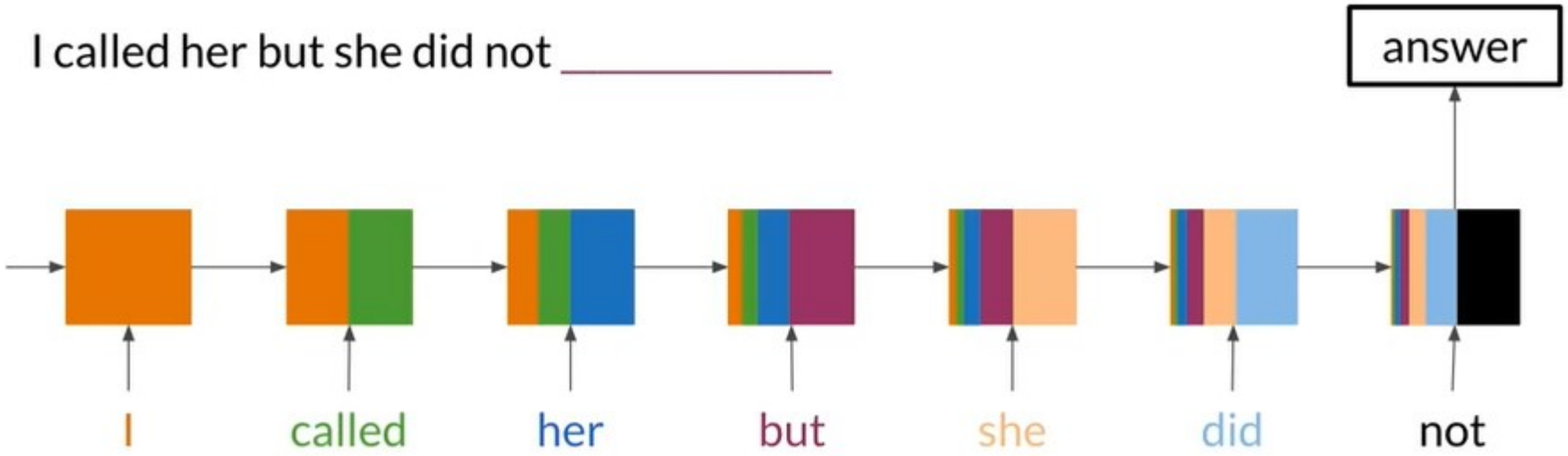
Word Representations

- Word representation techniques represent words as feature vectors.

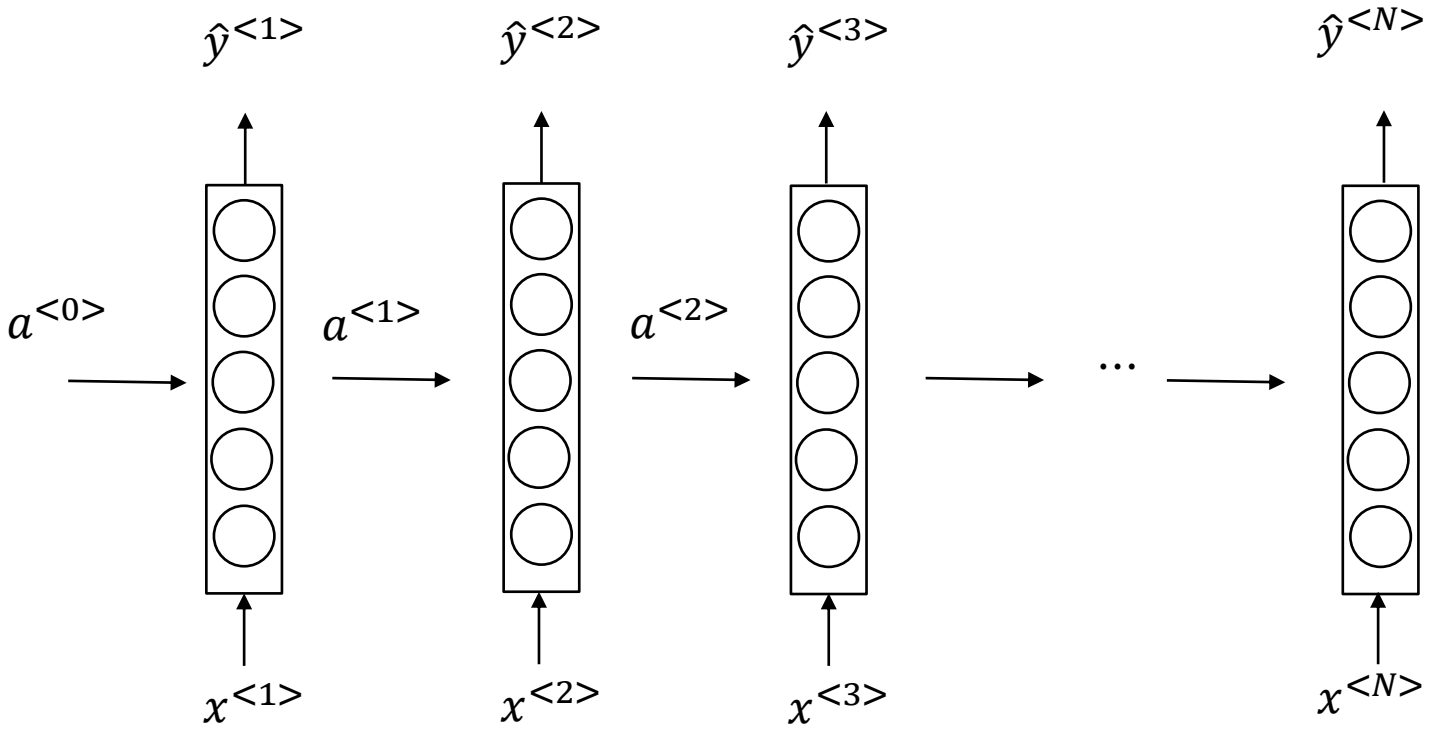


- Do you want to join me for some coffee?
- $x^{<1>}$ $x^{<2>}$ $x^{<3>}$ … … $x^{<8>}$

Recurrent Neural Networks

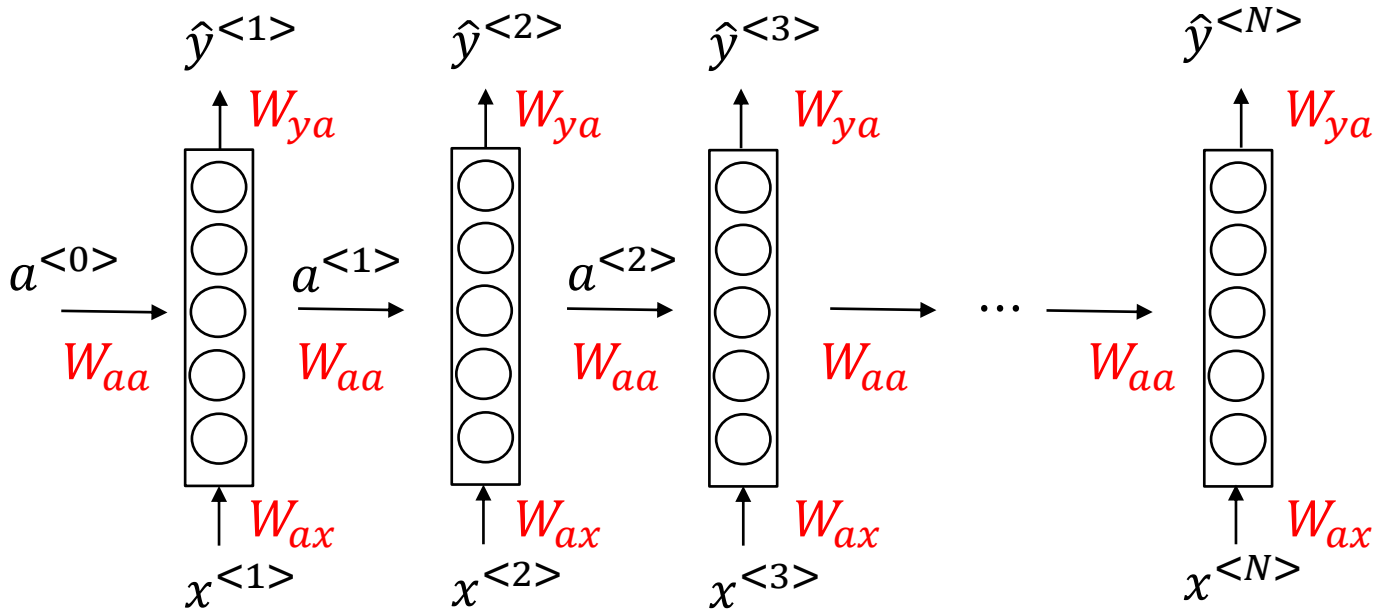


Recurrent Neural Networks



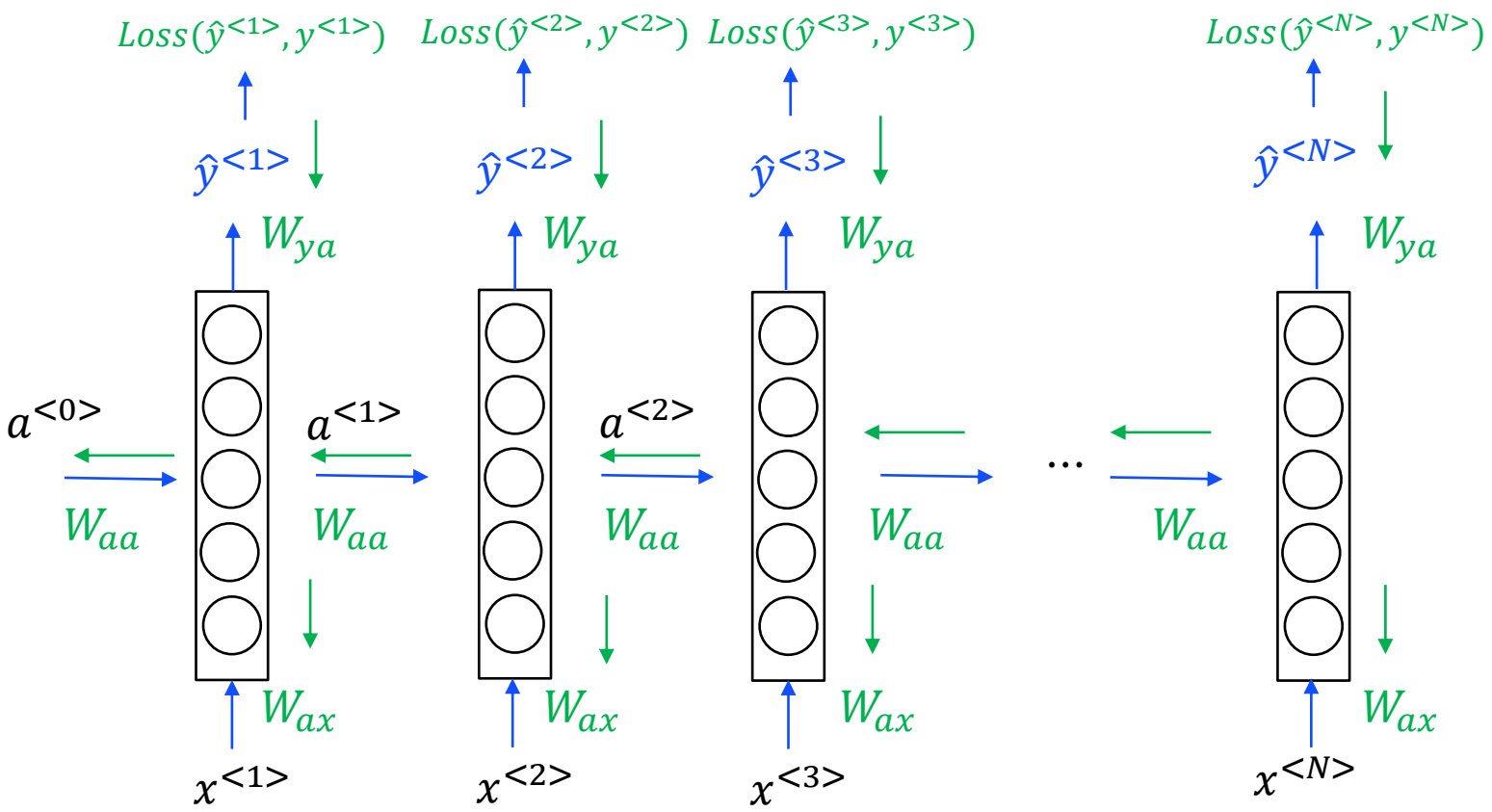
- The Recurrent Neural Networks can also be represented in compact (see the video).

Recurrent Neural Networks



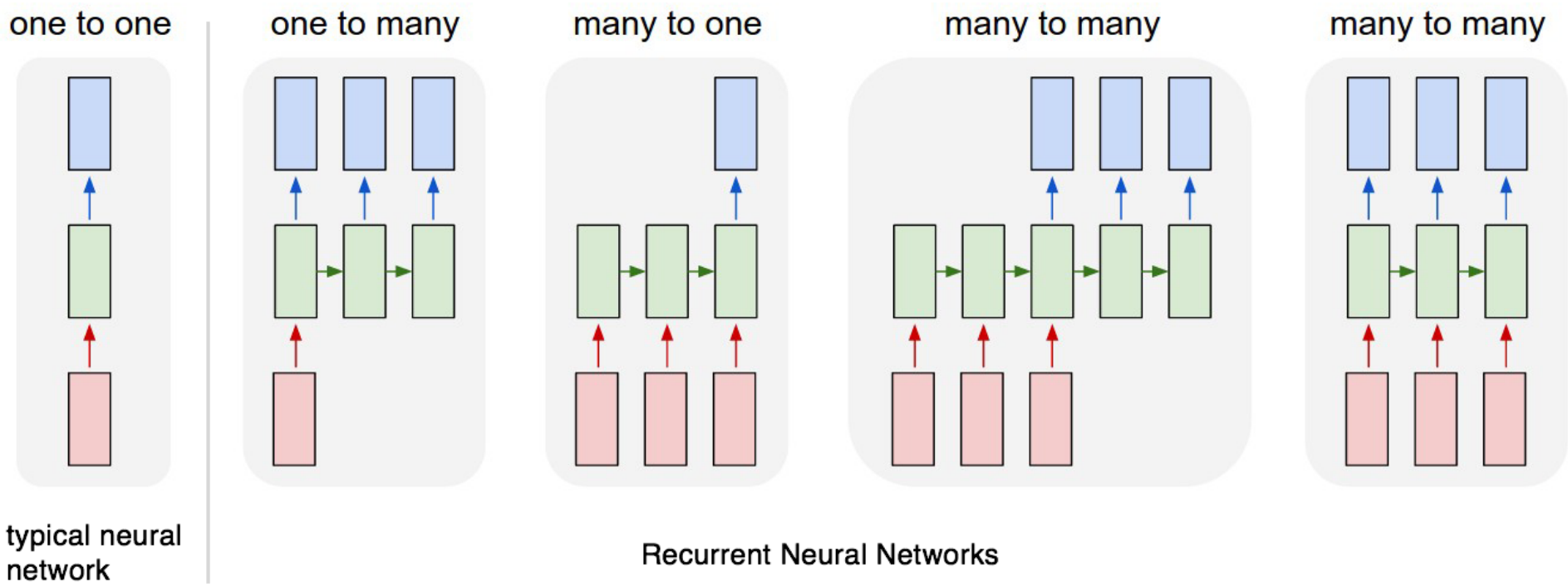
- Notice that all the W_{ax}'s are shared, all the W_{aa}'s are shared and all the W_{ya}'s are shared.
- The weight sharing properties makes out network suitable for variable-size inputs

Forward Propagation and Backpropagation

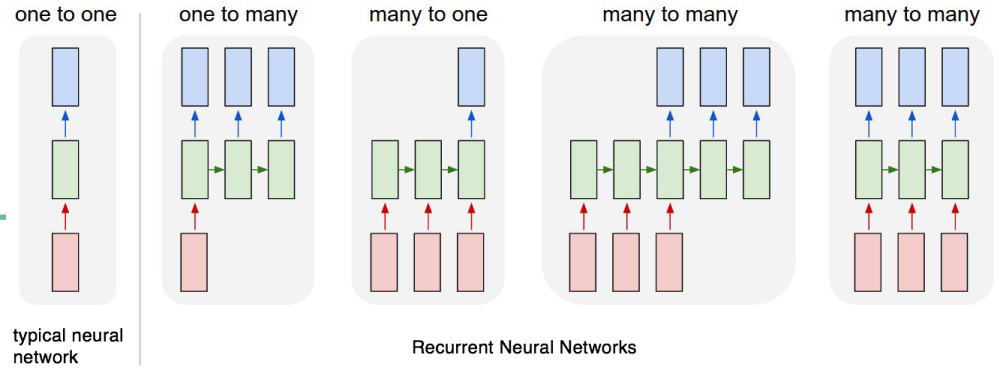


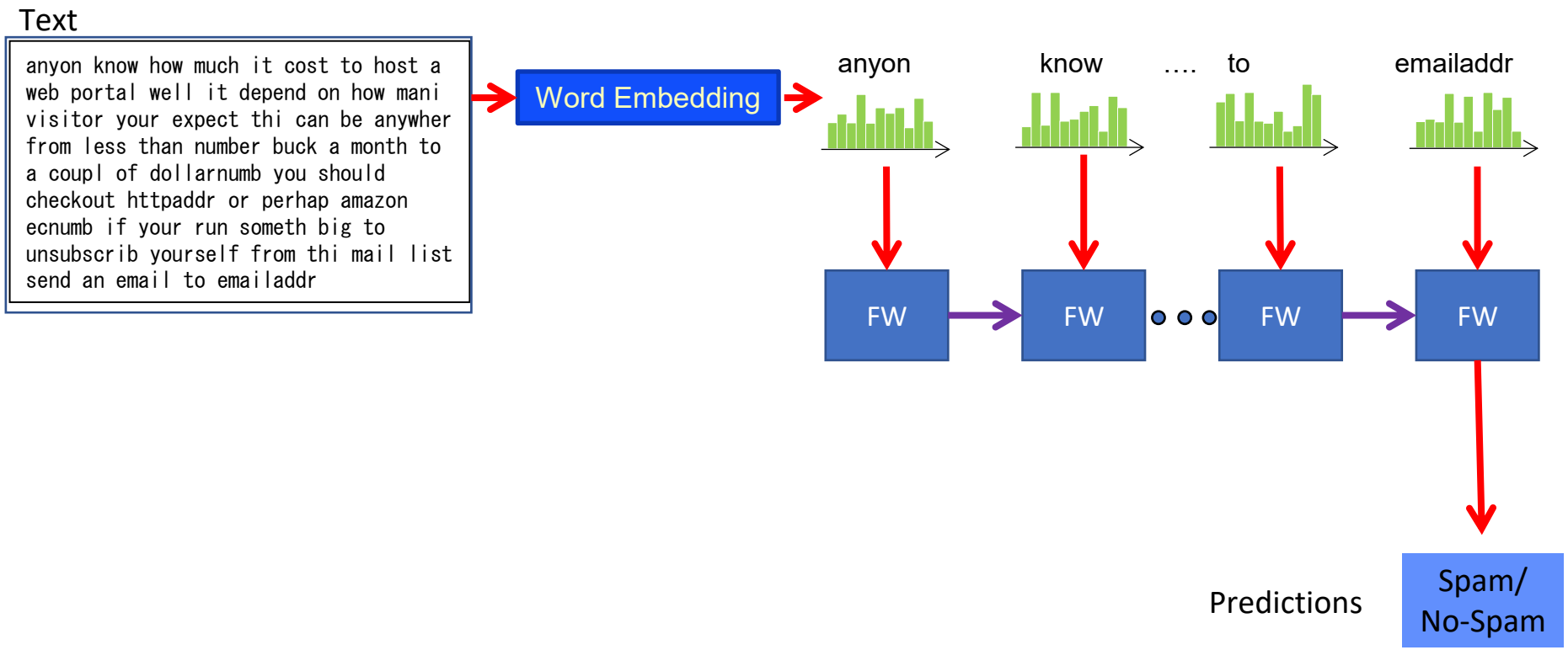
Backpropagation Through Time (BTT)

Diagram of different RNN sequence types

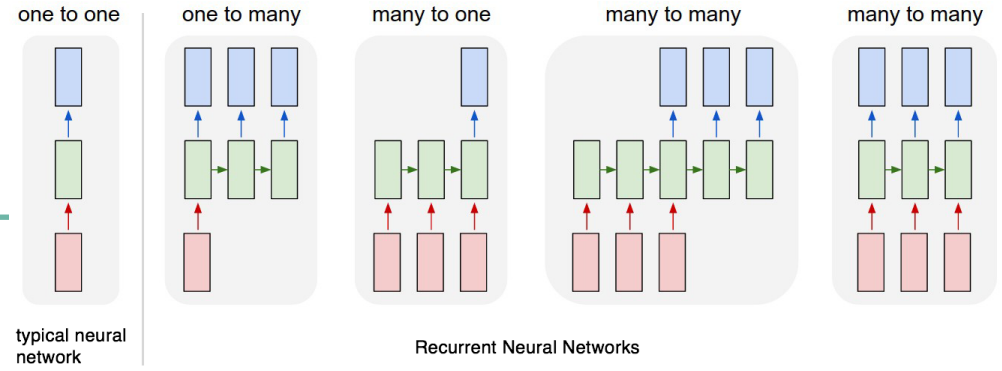


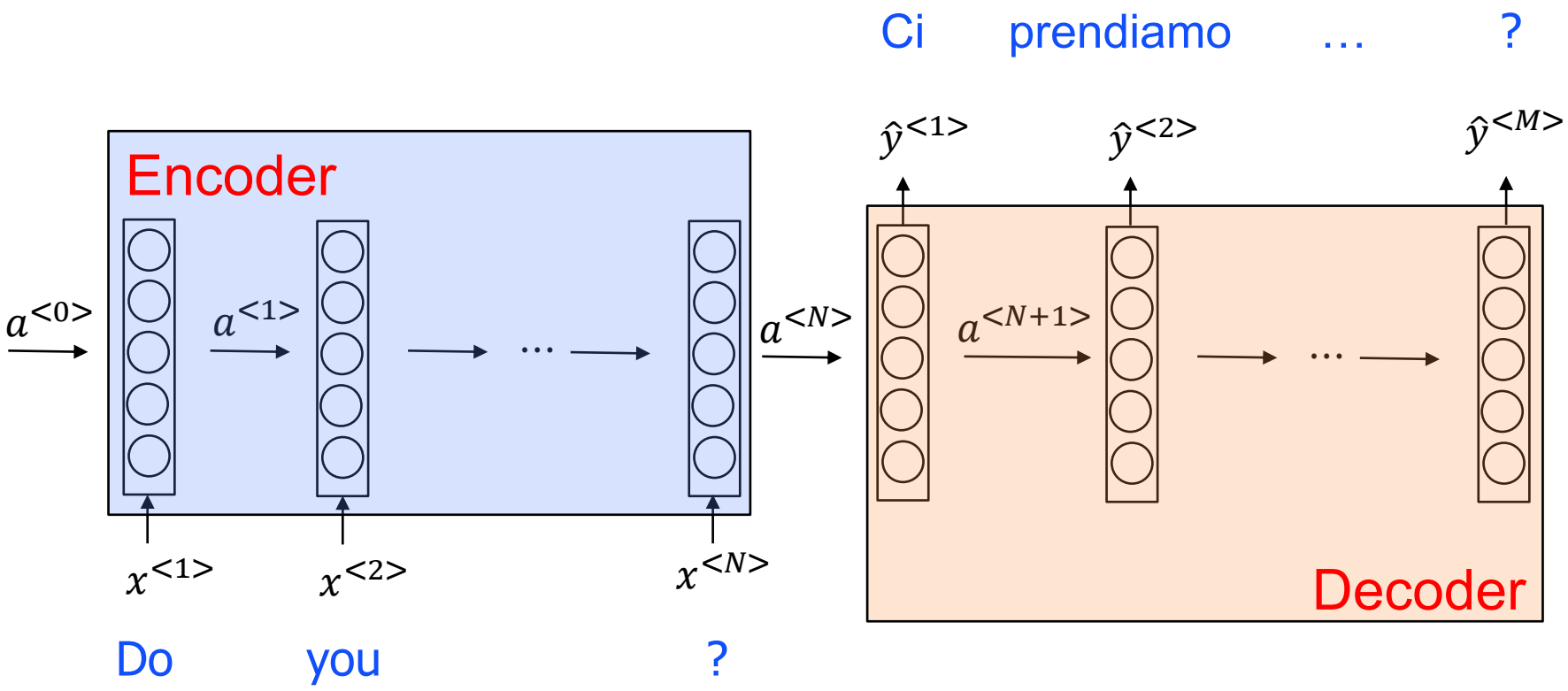
Text Classification using Word Embedding (Many to One)





Language Translation (Many to Many)





Automatic Named Entity Extraction (Many to Many)

- Locates and extracts predefined entities from text
- Places, organizations, names, times, and dates

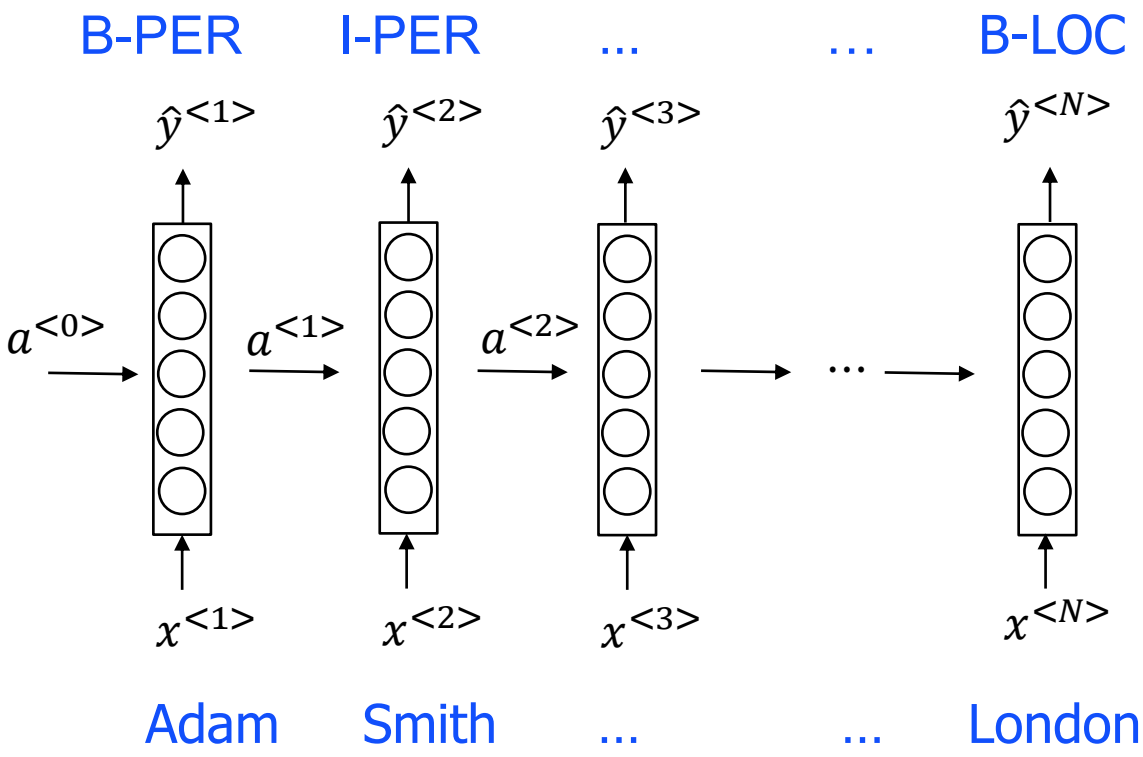
Automatic Named Entity Extraction (Many to Many)

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

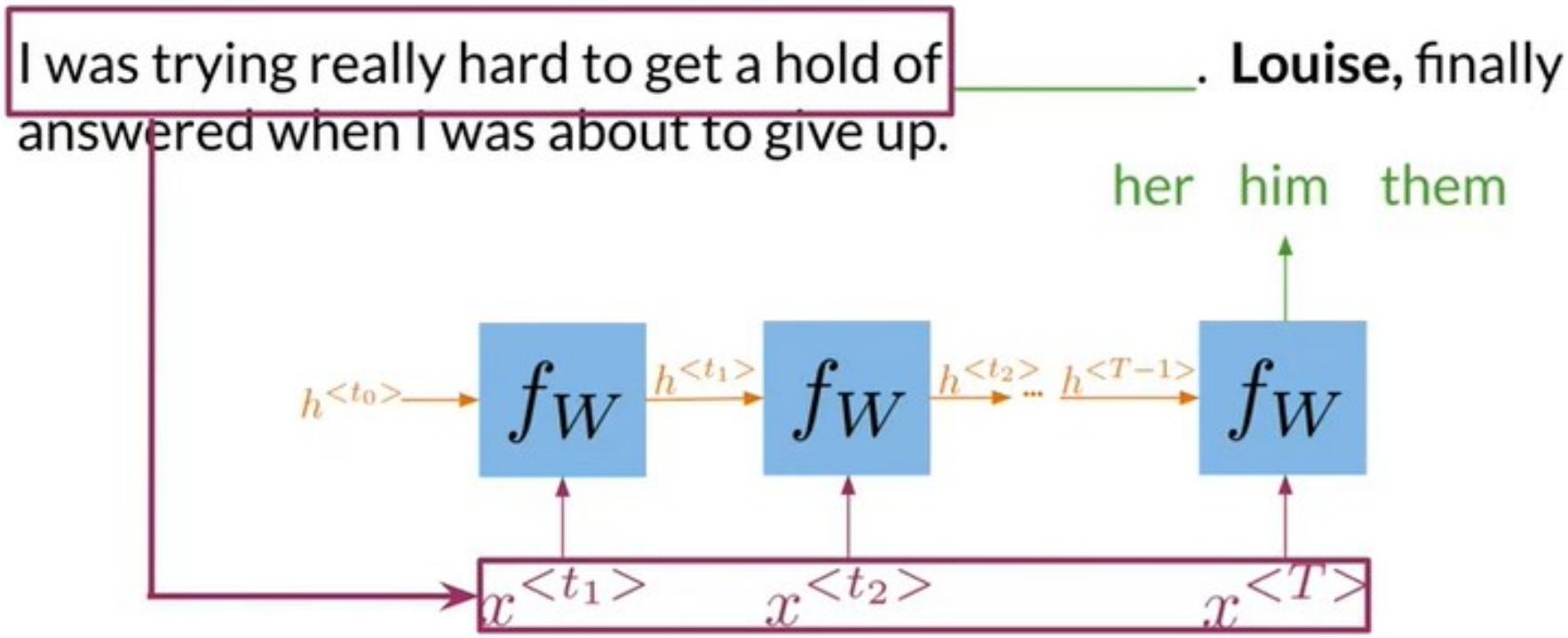
Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

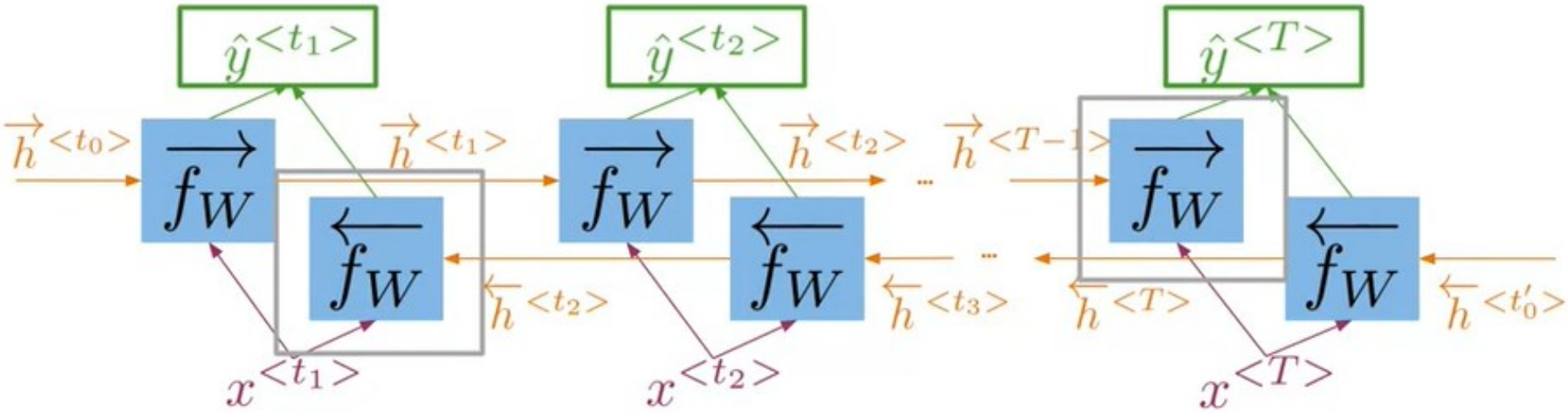
Automatic Named Entity Extraction (Many to Many)



RNNs

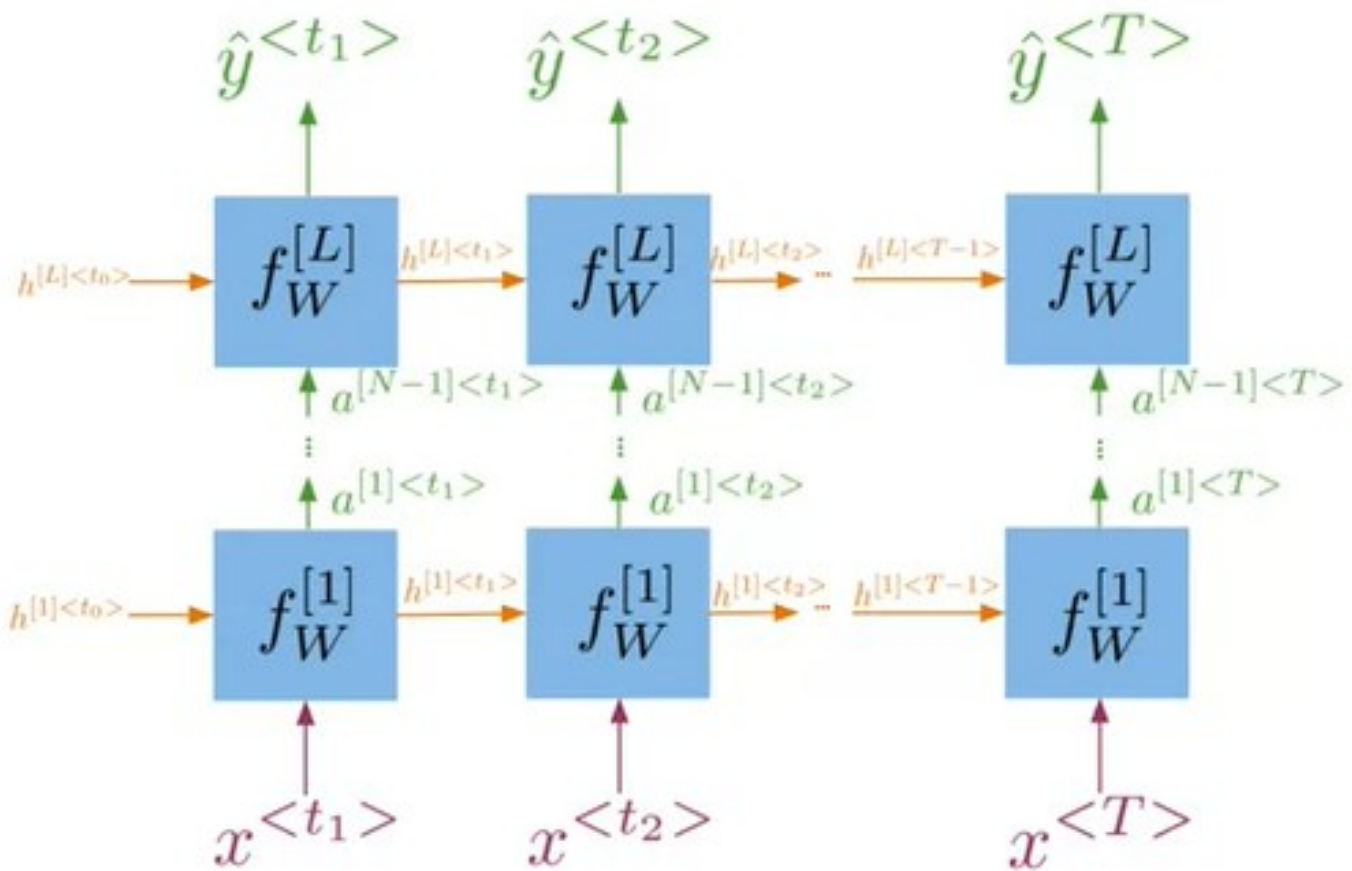

I was trying really hard to get a hold of
answered when I was about to give up.
. Louise, finally
her
him
them
$h^{<t_0>}$
f_W
$h^{<t_1>}$
f_W
$h^{<t_2>}$
...
$h^{<T-1>}$
f_W
$x^{<t_1>}$
$x^{<t_2>}$
$x^{<T>}$

Bi-Directional RNNs

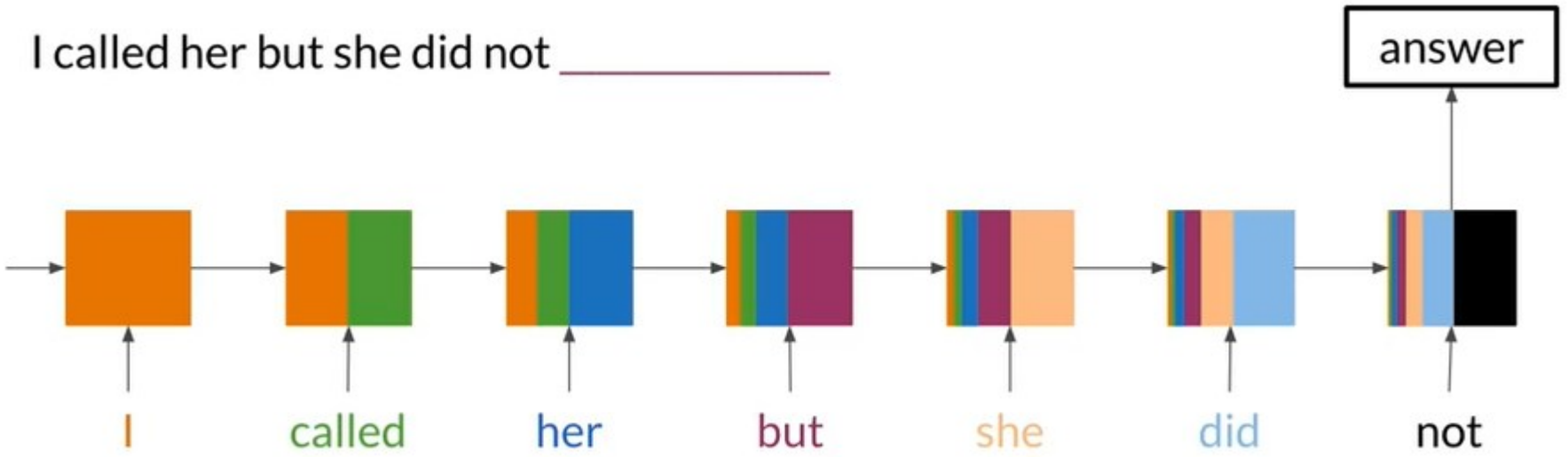


$$\hat{y}^{<t>} = g(W_y[\overrightarrow{h}^{<t>}, \overleftarrow{h}^{<t>}] + b_y)$$

Deep RNNs

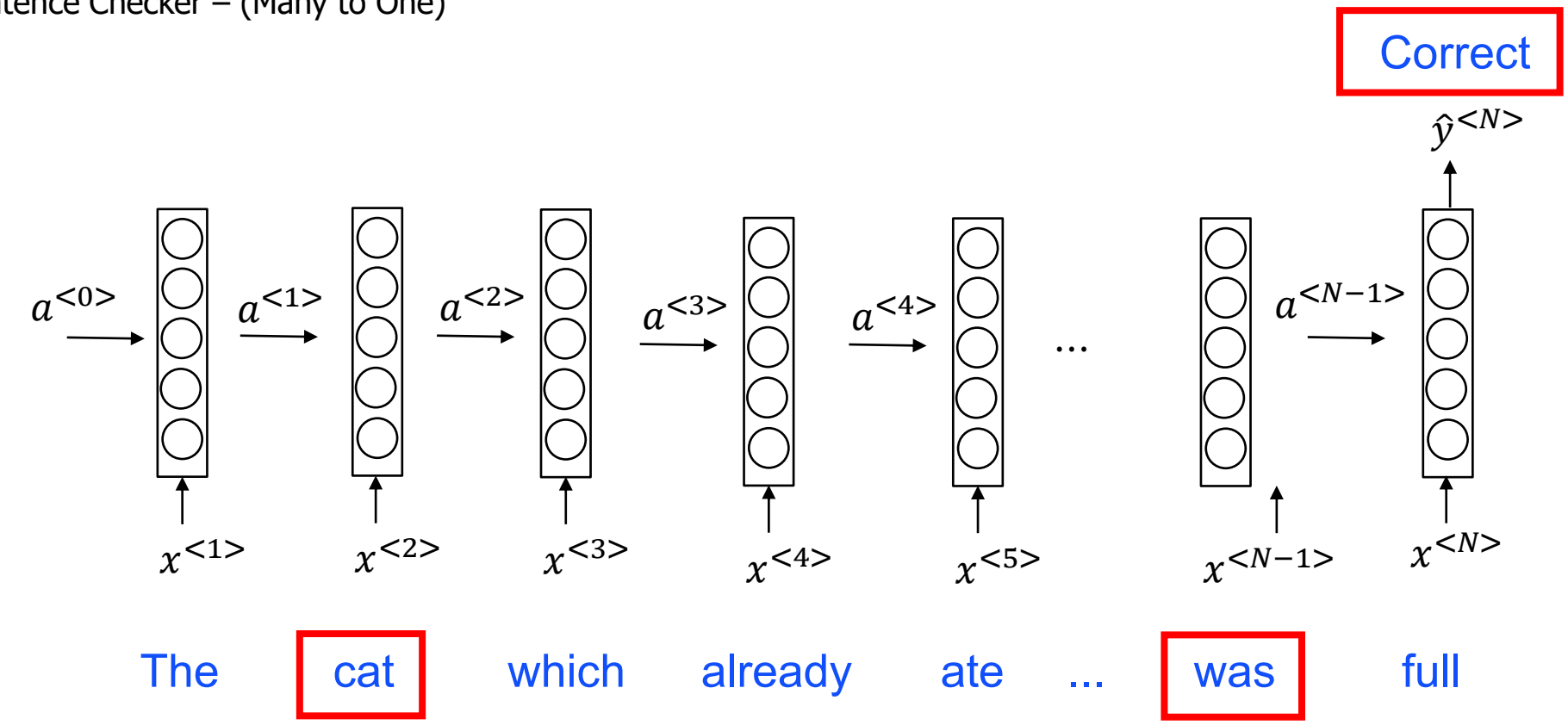

$\hat{y}^{<t_1>}$
$\hat{y}^{<t_2>}$
$\hat{y}^{<T>}$
$f_W^{[L]}$
$f_W^{[1]}$
$h^{[L]<t_0>}$
$h^{[L]<t_1>}$
$h^{[L]<t_2>}$
$h^{[L]<T-1>}$
$h^{[1]<t_0>}$
$h^{[1]<t_1>}$
$h^{[1]<t_2>}$
$h^{[1]<T-1>}$
$a^{[N-1]<t_1>}$
$a^{[N-1]<t_2>}$
$a^{[N-1]<T>}$
$a^{[1]<t_1>}$
$a^{[1]<t_2>}$
$a^{[1]<T>}$
$x^{<t_1>}$
$x^{<t_2>}$
$x^{<T>}$

Vanishing gradients with RNNs



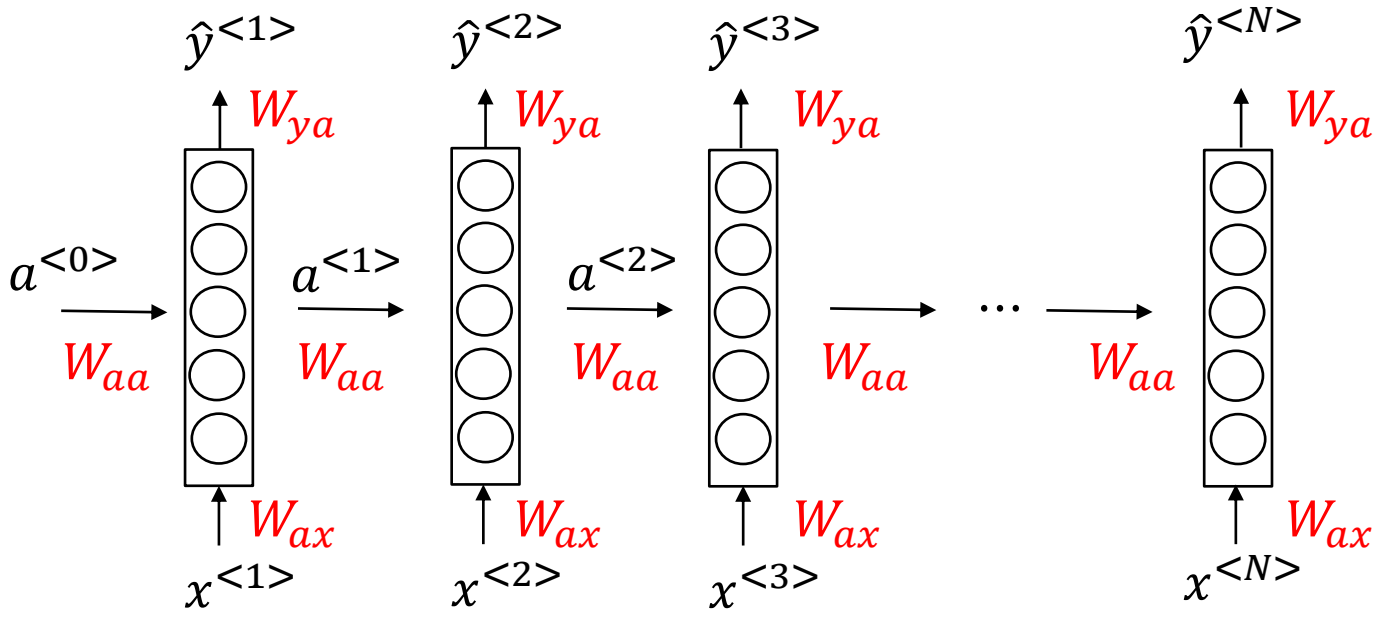
Vanishing gradients with RNNs

Sentence Checker – (Many to One)



RNN Equation

- $a^{<t>} = g_t(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$
- $\hat{y}^{<t>} = g'_t(W_{ya}a^{<t>} + b_y)$
- Where g_t is usually a ReLU/Tanh function and g'_t is usually a Sigmoid function

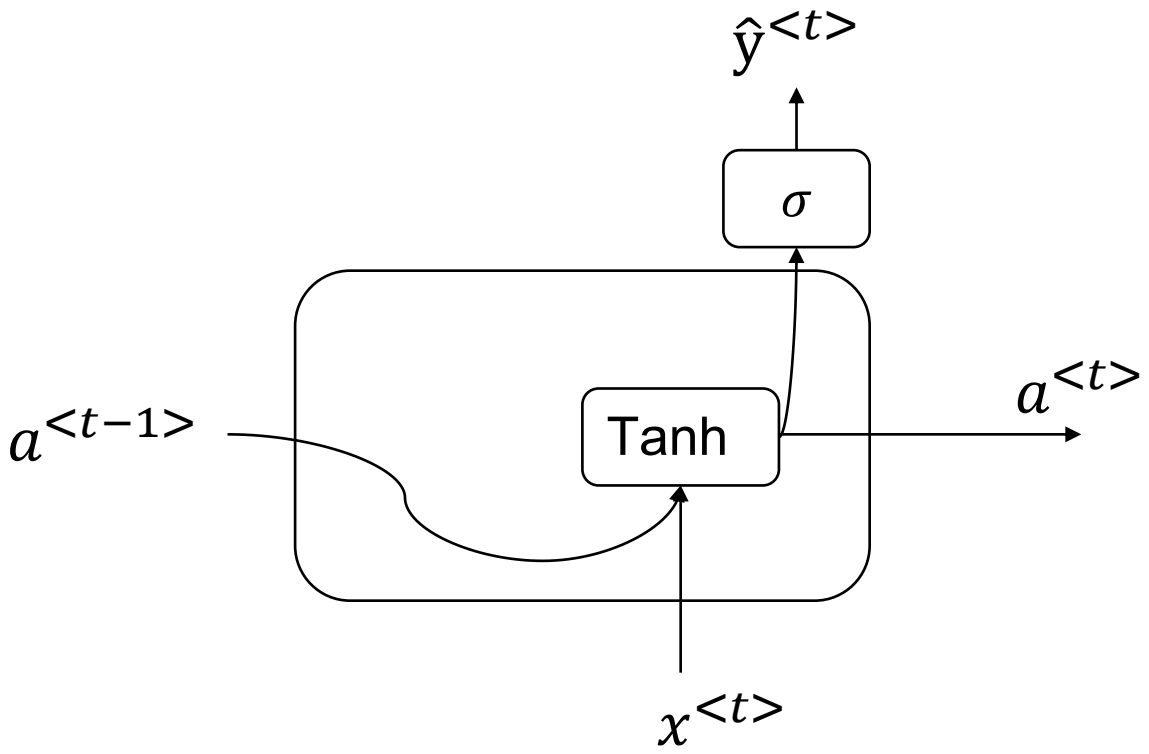


Simplified RNN Notation

- $a^{<t>} = g_t(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) = g_t(W_a[a^{<t-1>}, x^{<t>}] + b_a)$
- $\hat{y}^{<t>} = g'_t(W_{ya}a^{<t>} + b_y) = g'_t(W_y a^{<t>} + b_y)$
- Where g_t is usually a ReLU/Tanh function and g'_t is usually a Sigmoid function
- Where $W_a = [W_{aa}\ W_{ax}]$ and $[a^{<t-1>}, x^{<t>}] = \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix}$
- In fact,
- $[W_{aa}\ W_{ax}] \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} = W_{aa}a^{<t-1>} + W_{ax}x^{<t>}$

RNN Schematization: an example

- $a^{<t>} = tanh(W_a[a^{<t-1>}, x^{<t>}] + b_a)$
- $\hat{y}^{<t>} = \sigma(W_y a^{<t>} + b_y)$



Gated Recurrent Unit (GRU)

- Consider the example "The cat, which already ate …, was full"
- Let's define a **new variable c as Memory Cell**
  - c is a memory cell, which should remember that cat, in our example, is singular
- Let's define $c^{<t>}$ as the memory cell at the t time
- In this unit $c^{<t>} = \mathrm{a}^{<t>}$; we use different notation, because we will discuss about LSTM.

Gated Recurrent Unit (GRU) - (Simplified)

- Consider the example "The cate, which already ate …, was full"
- Let's define a candidate of c
  - $\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$
- Let's define a Gate u
  - $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$
  - Note $\Gamma_u \epsilon (0,1)$ and $\mathbf{u}$ **refer to "update"**
- Now, c is updated as following:
  - $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$

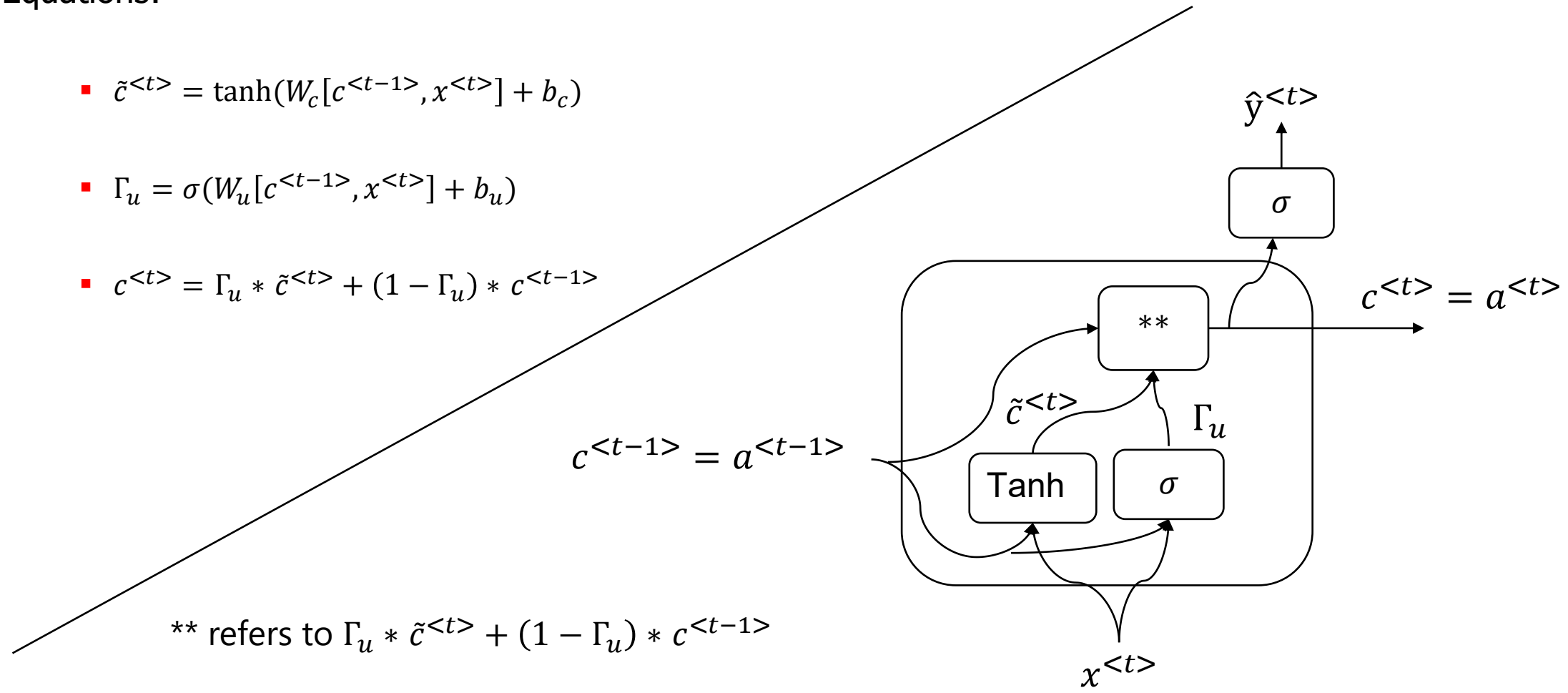
Gated Recurrent Unit (GRU) - (Simplified) - Intuition

- Equations:
  - $\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$
  - $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$
  - $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$
- Intuition of the idea of Gate Γ_u supposing $\Gamma_u \in \{0,1\}$

Γ_u=1	Γ_u=0	Γ_u=0	Γ_u=0	Γ_u=0	… …	Γ_u=1	Γ_u=0
c^0 =1	c^1=1	c^2=1	c^3=1	c^4 =1	… …	c^{N-1}=0.8	c^N =0.8
The	cat	which	already	ate	… …	was	full

GRU Schematization

- Equations:
  - $\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$
  - $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$
  - $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$



** refers to $\Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$

GRU (some more details)

- Equations:
  - $\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$
  - $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$
  - $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$
- $\tilde{c}^{<t>}$ can be a vector
- If, for example, $\tilde{c}^{<t>} \epsilon R^{100}$ than $\Gamma_u \ \epsilon R^{100}$ and $c^{<t>} \epsilon R^{100}$
- In addition, $*$ in the equation $c^{<t>}$ is an element-wise multiplication

FULL GRU

- Equations:
  - $\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$
  - $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$
  - $\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$
  - $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$

In the Gate Γ_r, $\boldsymbol{r}$ **stands for relevance**

- Why do we have also Gate r?

- Many researchers have experimented many different variations to address the vanishing gradient problems.
- Experiments in different contexts show this unit is robust and useful

GRU and Long Short-Term Memory (LSTM)

GRU

- $\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$
- $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$
- $\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$
- $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$
- $a^{<t>} = c^{<t>}$

LSTM

- $\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$
- $\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$
- $\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$
- $\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$
- $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$
- $a^{<t>} = \Gamma_o * c^{<t>}$

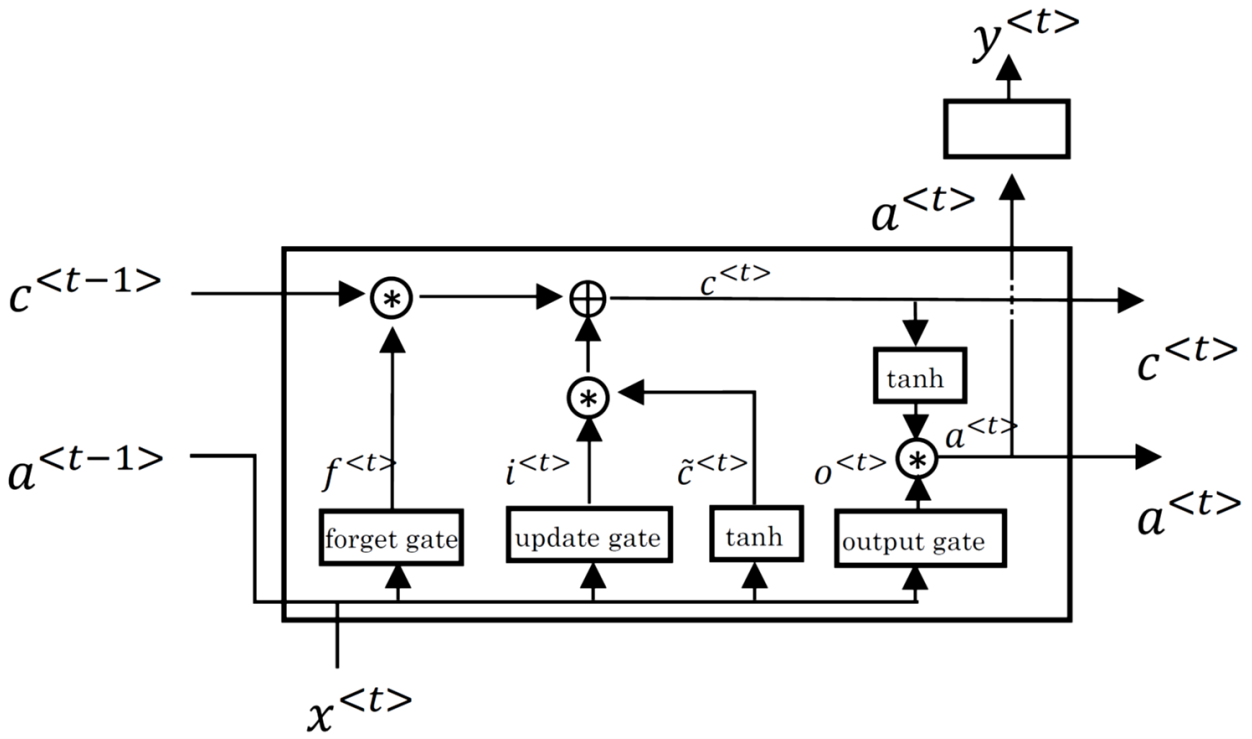
- Γ_u: Gate Update
- Γ_f: Gate Forget
- Γ_o: Gate Output

LSTM Schematization

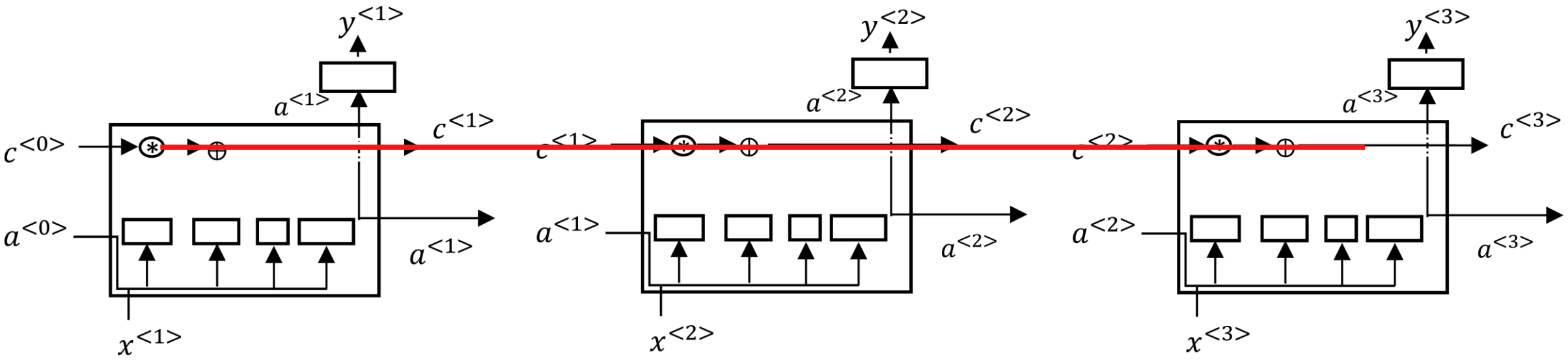
LSTM

- $\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$
- $\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$
- $\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$
- $\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$
- $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$
- $a^{<t>} = \Gamma_o * c^{<t>}$

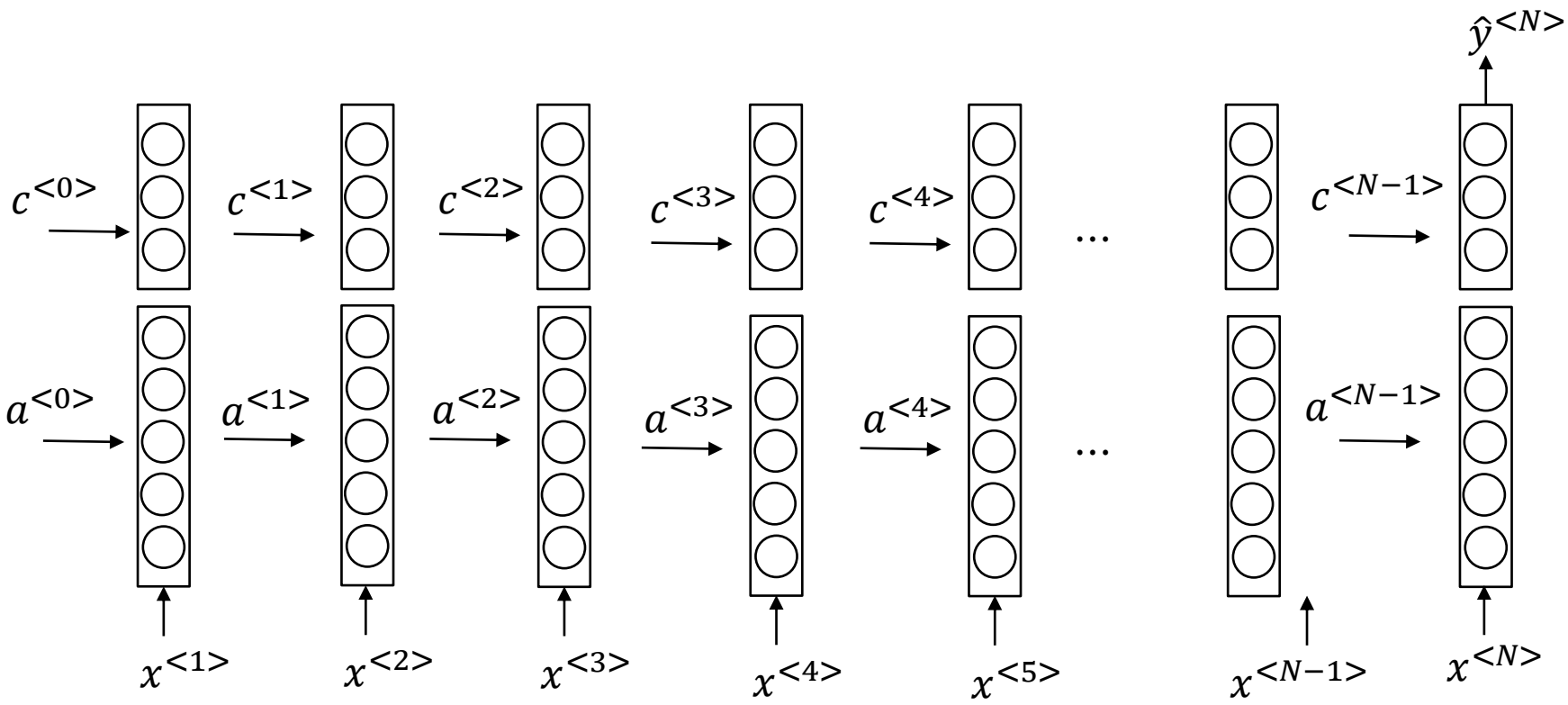
- Γ_u: Gate Update
- Γ_f: Gate Forget
- Γ_o: Gate Output



Sequence of LSTM



Long Short-Term Memory (LSTM) – Intuition



The C state change slowly through the time, so the LSTM architecture can better "remember" past features

Links:

Illustrated Guide to LSTM's and GRU's: A step by step explanation

The fall of RNN / LSTM

Understanding LSTM and its diagrams